Context-Dependent Automatic Response Generation Using Statistical Machine Translation Techniques

Andrew Y. Shin Ryohei Sasano Hiroya Takamura Manabu Okumura
Tokyo Institute of Technology
shin@lr.pi.titech.ac.jp {sasano,takamura,oku}@pi.titech.ac.jp

Abstract

Developing a system that can automatically respond to a user's utterance has recently become a topic of research in natural language processing. However, most works in the field take into account only a single previous utterance to generate a response. Recent works demonstrate that the application of statistical machine translation (SMT) towards monolingual dialogue setting has a great potential, and we exploit the approach to explore the context-dependent response generation task. We attempt to extract relevant and significant information from the wider contextual scope of the conversation and incorporate it into the SMT techniques. We also discuss the advantages and limits of this approach through our experimental results.

1 Introduction

With the advent of Siri and Google Now, conversational agent systems have become a ubiquitous application that is not only useful for entertainment or task-specific purposes but also as a general user-assistance application. While such commercial systems have proven to be a useful tool, they are not without drawbacks. Most systems generate a response solely based on the single previous utterance, failing to take into account the overall context of the conversation, or significant information from the past utterances.

Some recent works have approached the task from a different angle with the application of SMT techniques into the task, often with various goals and moderate success (Ritter et al. 2011, Hasegawa et al. 2013). However, such attempts also fail to take into account the overall context throughout the conversation, and often display unreliable alignments between source and target sentences, especially when the source sentences are shorter than target sentences.

We attempt to implement a context-dependent model where we try to balance the alignment by adding the semantically critical words from previous utterances to the most recent statement. By doing so, we hope not only to generate more reliable alignments, but also to be able to take semantics from broader scope of the conversation into account.

2 Response Generation using SMT

To our knowledge, Ritter et al. (2011) made the first attempt to apply SMT techniques into monolingual dialogue setting for the response generation task, and our baseline model follows the approach described in the paper. In SMT, a string f in a source language is translated into a string e in a target language according to probability distribution $p(e|f)$. However, the paper remarks the frequently observed structural resemblances in stimulus-response pairs of the same language, as in the following example, and treats the response as the *translation* of the stimulus.

Stimulus: What is your name?
Response: My name is Andrew.

Of course, there are far more cases where syntactic structures between stimuli and responses differ greatly, but the results show that their high semantic correlation frequently leads to desired outputs.

However, the application of SMT into the response generation task has some fundamental problems. First, it cannot take into account what was previously discussed in the conversation. If the most recent statement brings a completely new topic, or it has sufficient information in itself, then such problem is obscured. In many cases, however, the problem is apparent; for example, when the source statement is too short, as shown in the following example. The response generated by the baseline model to the most recent stimulus in this example is “that,” which only mimics the syntactic structure but fails to deliver any meaningful content.

(example):

A: Is something going on today?
B: Of course, it's dad's birthday.
A (most recent stimulus) **:** What?!
B (target) **:** Oh my, you really didn't know?

Another problem is with the alignment. Conversational setting obviously does not guarantee that a source sentence and a target sentence are of similar lengths. As in the example above, unbalanced lengths between source and target sentences cause many words to be aligned to an empty word. As a result, neither Hidden Markov Model nor IBM Model can stably handle the alignment between unbalanced source and target sentences in conversational setting.

3 Context-Dependent Model

We now describe the model we worked on in order to deal with the problems discussed in the previous section. We observed how source and target sentences of different lengths result in poor alignments. We thus want to minimize the gap in the sentence lengths and make them as equal as possible, while hopefully retaining some of the core semantics from the earlier parts of the conversation. Obviously we cannot align multiple source sentences with a single target sentence, but we can use some semantically relevant tokens to fill in the gap. The question is how to determine whether a token is of high relevance to the topic throughout the conversation. We rely on the Fisher's exact test for the task, which shows strong performance even when the counts of words are small.

3.1 Fisher's Exact Test

For each pair of sentences consisting of sentence *Si* and the following sentence *Si+1*, we compute the *p*-value from the Fisher's exact test for all possible pairs of words $s \in Si$ and $t \in Si+1$. Fisher's exact test uses hypergeometric distribution to compute the exact probability of a particular joint frequency:

$$\frac{C(s)!\,C(\neg s)!\,C(t)!\,C(\neg t)!}{N!\,C(s,t)!\,C(\neg s,t)!\,C(s,\neg t)!\,C(\neg s,\neg t)!}.$$

In the equation above, *C* denotes the counts of the word in the parenthesis throughout the training data, and *N* denotes the sum of counts for all words. If *p*-value of the test is below the threshold, we add the words to the input sentence in the same order they appeared in the conversation, avoiding duplications. We experimented with different *p*-values, and observed that sentence lengths are most balanced when p-value is .0001. Applying this method to the stimulus in the example conversation from the previous section results in the following new stimulus. Parenthesis indicates the newly added words from the earlier utterances:

A: (today birthday) What?!

As shown in the equation, Fisher's exact test involves many factorials, and thus has traditionally been considered inappropriate for large sample size, but Moore (2004) presents a workaround to make the computation feasible (Ritter et al. 2011). We computed the test with fast, freely available SciPy module (Jones et al. 2001).

3.2 Discussion

This method clearly breaks down the grammatical integrity of the input sentence, and the alignments will thus have much higher focus on semantic co-occurrences of tokens rather than structural resemblance of input/output. For that reason, we opted not to use the reordering table for our model.

By adding the words, we risk confusing the translation model and losing grammaticality of the output. However, we are convinced that it is not to the extent where its effects seemingly degrade the performance.

First, since we opted not to use the reordering table, we have higher reliance on the language model for grammaticality. This does not affect the grammaticality of the output because the language model is constructed upon the target language only, in this case the responses to the stimuli, which are at least supposed to be grammatically correct, albeit with frequent violations due to the nature of SNS.

Second, the newly added tokens are of high relevance to each other, so the new input sentence with added tokens frequently demonstrates fair semantic coherence, despite ungrammaticality. Such semantic relevance is often valid when coupled with the target sentence as well, so counting the co-occurrences of tokens from earlier sentences and tokens in the target string can strengthen the semantic ties, which the example in the previous section lacked.

4 Data

Our model was implemented using Moses toolkit with KenLM as the language model. We built our training, tuning, and test data set from Twitter. First, we collected tweets whose '*in_reply_to_ID*' field was not '*Null*,' i.e., they had a "parent" tweet to whom it replied. We then retrieved the tweets that correspond to '*in_reply_to_ID*' field of the tweets collected in the first step, checked their '*in_reply_to_ID*' field, and recursively collected the parent tweets that correspond to it.

Further restrictions were necessary. First, we restricted each conversation to have between 3 and 10 utterances. Also, in Twitter, it is possible to post a reply to one's own tweet. We concluded that this was against the nature of our research, and filtered the collected data under a requirement that each consecutive tweets must be from different users; in other words, speakers in the tweets have to take turns.

Additionally, we dealt only with alphabetical characters, filtering out all non-alphabetical characters including numerics. Similarly, certain expressions that are unique to Twitter have also been filtered out; for example, *@username* and *#hashtag*.

Our training data consist of 909,650 pairs of sentences for both models, with roughly 9.4M words for the baseline model and 11.7M words for our model. Table 1 shows the average sentence length of training data for both baseline and context-dependent models, each again divided into source and target parts. Note how the average sentence length increases in the source language for our context-dependent model, while it is identical to that of the baseline model for the target language.

Model	Source	Target
Baseline	10.41	10.28
Context-Dependent	12.87	10.28

Table 1: Average Sentence Lengths for Training Data

5 Evaluation

5.1. Preliminary Experiment

Our model is supposed to help in cases where a source sentence is shorter than a target sentence. It should be noted, however, that the relative length of a source sentence in comparison to a target sentence cannot be a criteria for selective application of our model, since we will not be able to know what the target response to our source sentence will be in a real testing environment. Thus, we considered it more practical to determine a length *n* of the source sentence that can be a criterion for whether to apply our model.

We first checked whether there was a relation between the source length and the target length. We took 10 samples with each consisting of 10,000 pairs of the source sentence length and the corresponding target sentence length, and computed their regression. In spite of sparse distribution and low R^2, the regression shows consistent tendency to converge roughly at y=0.3x + 7, with y being the length of target sentences, and x the length of source sentences. This is consistent with the average sentence length of the training data, and shows that source sentences are shorter than target sentences on average when the length of source sentences is less than 10.

We performed a preliminary experiment to confirm the criterion length. We requested a human intelligence task (HIT) on Amazon Mechanical Turk with 200 multiple-choice type questions. Each question showed a part of conversation randomly chosen from our test data, and the workers were required to choose the better response from two possible choices, corresponding to responses generated by the baseline and our model. Of the 200 questions, 104 questions had the last statements' lengths shorter than 10, 11 questions equal to 10, and 85 questions greater than 10.

Figure 1 shows the relation between the source sentence length and the selection rate of our context-dependent model. Up to |S|=10, there clearly is an inversely correlated relation, and beyond 10 is without any notable pattern. In fact, R^2 up to |S|=10 is 0.96, which shows a strong correlation.

5.2. Final Evaluation

One of the challenging aspects of the researches on conversation is its distinct nature in which there is an extremely wide range of acceptable candidates to a stimulus, unlike usual bilingual translation tasks where there are typically pre-set candidates to be referenced with high reliability. For that reason, the usual automatic evaluation metrics for SMT cannot be used as the primary way to evaluate the experimental results in conversational setting, and we resort to human manual evaluation as our primary source of evaluation.

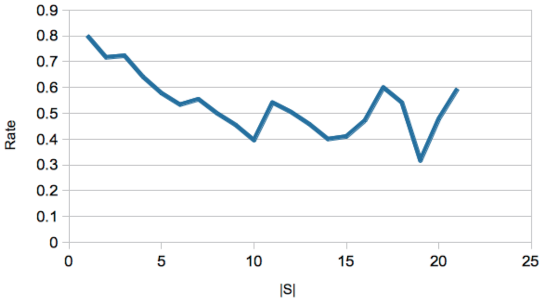

Figure 1: Source Sentence Length vs. Selection Rate of Context-Dependent Model

We performed a human evaluation on Amazon Mechanical Turk for source sentence lengths limited to less than or equal to 10. The evaluation task consisted of 200 questions and was done by 20 workers. Each question was a ranking task in which workers were given a part of conversation and were required to rank the responses that followed the conversations. For all questions, workers were given three responses; the actual response on Twitter, one generated by the baseline model, and one by our context-dependent model.

Table 2 compares our model's performance against the baseline model and the actual responses. Fraction is the number of choices that ranked our model higher divided by the total number of choices made. Mutual agreement is calculated by S coefficient. In our case, the mutual agreement is in the range of 0.4 to 0.6, which corresponds to moderate agreement. Note that fraction and mutual agreement are computed from the total number of *choices* made by the workers, whereas binomial *p*-value is calculated with the number of *questions* for which our model performed better, since each question is independent of the others. Overall, the table shows that our model was preferred over the baseline model, but performed poorly against the actual responses as expected. Table 3 shows the fraction of each model for each ranking and their average rankings. Our model outperforms the baseline model in higher rankings. Table 4 features examples of responses generated by each model and the actual responses on Twitter, along with their average ranking in the final evaluation. In all examples shown, our model was ranked higher than the baseline model, and the last example shows a case where our model was ranked higher than the actual response.

Conversation	Response*	Rank
A: like youre talking about the stupidest things ever. its annoying. **B:** who is this about	**1:** the ppl behind you	1.15
	2: what are they talking about	2.85
	3: like i said im takling seriously are you	2
A: It's nearly 11 at night and its still warm **B:** where do you live? **A:** uk	**1:** oh that makes sense	1.3
	2: london	2.65
	3: i live in london	2.05
A: Making the most of working in this weather **B:** keeping busy then, yeh? Time to lean, time to clean...	**1:** exactly what Graeme Fergie said although he was so jealous of our invention	1.95
	2: so yeah i like into you	2.35
	3: I am working to build it taste like straight	1.7

*1 refers to the actual response on Twitter, 2 is a response generated by the baseline model, and 3 is by context-dependent model.

Table 4: Examples of Responses

vs. Model	Fraction	Binomial p	Agreement
Baseline	.555	2.8e-03	.497
Actual	.236	1.7e-24	.556

Table 2: Comparison with Baseline Model and Actual Responses

Model	Rank 1	Rank 2	Rank 3	Avg. Rank
Actual	.686	.175	.138	1.45
Baseline	.137	.389	.551	2.34
Proposed	.177	.436	.387	2.20

Table 3: Comparison of Models' Performance on Human Evaluation

6 Conclusion & Future Work

As we observed in the experimental results, our context-dependent model can outperform the baseline model in a wider scope of conversations. Although it performs poorly against the actual responses, it can occasionally outperform them especially when the actual responses divert from the topic, or have poor coherence and grammaticality.

One of the improvements is likely to come from attempting different methods to extract the core tokens from the past utterances. We consistently relied on the Fisher's exact test throughout the research, but other approaches may perform better. Alternatively, we may try different weight systems depending on the distance between the past utterances and the current utterance. This follows the intuition that the more previous the utterances are, the less relevant they tend to be to the current topic. Another weight may be given to tokens depending on whether it is from the same speaker as the current utterance or a different speaker, since it would generally make more sense for a particular speaker not to repeat him/herself.

References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* Vol. 16, pages 79-85.

Takayuki Hasegawa, Nobuhiro Kaji, and Naoki Yoshinaga. 2013. Predicting and Eliciting Addressee's Emotion in Online Dialogue. In *ACL*, pages 964-972.

Eric Jones, Travis Oliphant, and Pearu Peterson et al. 2001. SciPy: Open Source Scientific Tools for Python. *http://www.scipy.org/*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *HTL/NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL*, pages 177-180.

Daniel Marcu, and William Wong. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *EMNLP*.

Richard C. Moore. 2004. On Log-Likelihood-Ratios and the Significance of Rare Events. In *EMNLP*.

Franz J. Och, and Hermann Ney. 2000. Improved Statistical Alignment Models. In *ACL*.

Franz J. Och, and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *COLING*, Vol. 2, pages 1086-1890.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *EMNLP*, pages 583-593.